# 卷积网络的进化 – 更小？更快？更强？

在上一章中，我们已经动手实现了一次LeNet。LeNet是当代卷积神经网络的雏形，它由Yann LeCun等人在1998年提出。LeNet是最早被广泛应用于手写数字识别任务的神经网络之一，它在推动计算机视觉和深度学习领域的发展方面起到了重要的作用。不过，直到2012年AlexNet的提出，CNN才开始连同深度学习一起火热起来。在这一章中，我们将介绍卷积网络的进化史，还将讨论卷积核的尺寸设置对CNN性能的影响。

## AlexNet

AlexNet是一种经典的卷积神经网络（Convolutional Neural Network，CNN），由Alex Krizhevsky、Ilya Sutskever和Geoffrey Hinton于2012年提出。它在ImageNet图像分类挑战赛（ImageNet Large Scale Visual Recognition Challenge，简称ILSVRC）上取得了显著的突破，标志着深度学习在计算机视觉领域的崛起。

AlexNet采用了深层的卷积神经网络架构，在当时是一种创新和颠覆性的设计。它由五个卷积层和三个全连接层组成。在卷积层之间使用了非线性激活函数ReLU（Rectified Linear Unit），并采用了局部响应归一化（Local Response Normalization, LRN）进行正则化。此外，AlexNet还使用了Dropout技术来减少过拟合。

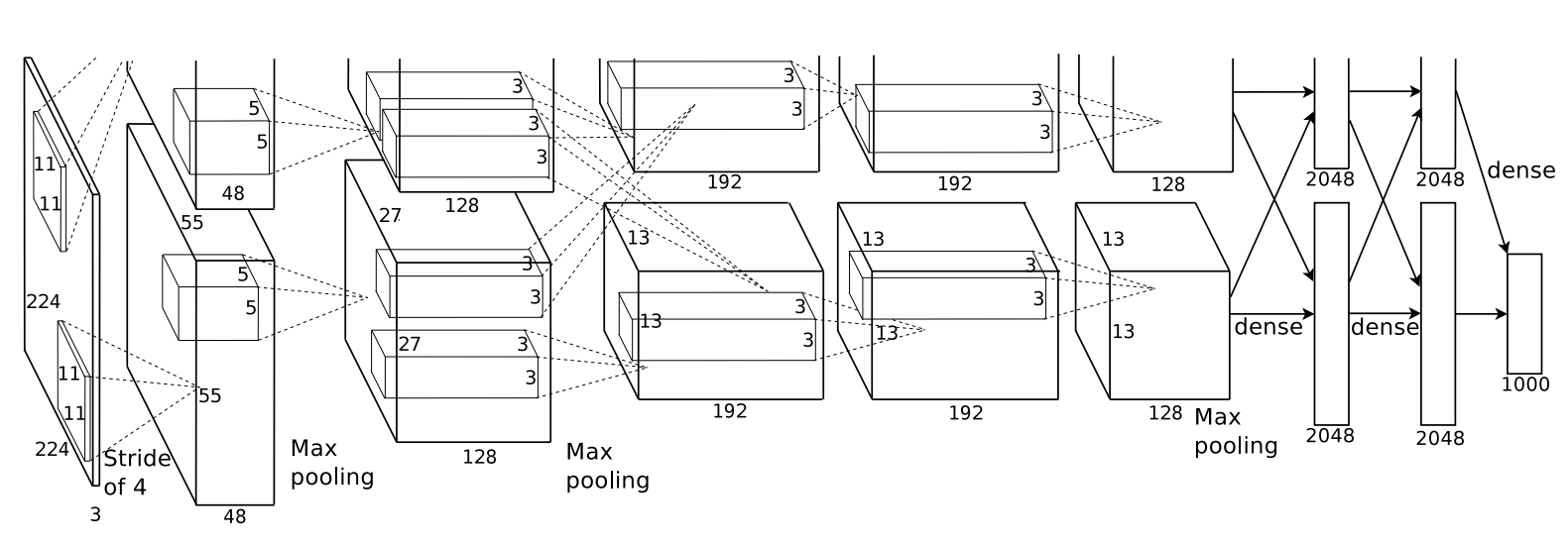


图 4‑1 AlexNet结构图[1]

图4-1展示了AlexNet的网络结构图。AlexNet分为上下两路，上半看起来不完整的部分是与下半部分完全相同的。在AlexNet产生的年代，如果将两路合并起来，由于当时的硬件性能约束，单个GPU将无法进行网络训练。于是，AlexNet采用了分为两路的技巧，让两个部分分别通过两个GPU并行训练。

|  |
| --- |
| 结构 |
| CONV1 |
| MAXPOOL1 |
| NORM1 |
| CONV2 |
| MAXPOOL2 |
| NORM2 |
| CONV3 |
| CONV4 |
| CONV5 |
| MAXPOOL3 |
| FC6 |
| FC7 |
| FC8 |

表 4‑1 AlexNet结构

AlexNet设置了5个卷积层。除了首个卷积层使用了11x11的大卷积核外，余下的四个卷积层使用的都是尺寸较小的卷积核。第二个卷积层使用了5x5的卷积核，而余下三个卷积层使用的卷积核都是3x3的。

|  |  |
| --- | --- |
| 卷积层 | 卷积核大小 |
| CONV1 | 11x11 |
| CONV2 | 5x5 |
| CONV3 | 3x3 |
| CONV4 | 3x3 |
| CONV5 | 3x3 |

表 4‑2 AlexNet卷积层

AlexNet还使用了Dropout的技术。在训练过程中，Dropout 随机将一部分神经元的输出置为零（即失活），以一定的概率保留其他神经元的输出。这样做的效果是，在每次训练迭代中，都会有不同的神经元被失活，因此网络无法过度依赖某些特定的神经元，从而减少了神经元之间的耦合性，增加了网络的鲁棒性。

在2012年的比赛中，AlexNet的突破性表现使其大幅度领先于其他参赛模型，大大提升了图像分类的准确率。

AlexNet在深度学习和计算机视觉领域具有重要的学术意义。它首次将深度卷积神经网络引入计算机视觉任务，并证明了通过深层网络的使用可以提高图像分类的准确率。AlexNet的成功促进了深度学习的研究和应用，并激发了更多复杂神经网络架构的开发，如VGG、GoogLeNet和ResNet等。

此外，AlexNet的设计也对深度学习研究方法产生了重要影响，如ReLU激活函数、局部响应归一化和Dropout等技术的引入，为深层神经网络的训练和优化提供了指导和启示。

总的来说，AlexNet的出现标志着深度学习在计算机视觉领域的突破，对于深度学习的发展和推广产生了深远影响。

## GoogLeNet和Inception

GoogLeNet是指Google团队在2014年提出的深度卷积神经网络架构，也被称为Inception-v1。随后，为了进一步改进网络结构和性能，Google团队陆续提出了一系列基于Inception的网络版本，分别是Inception-v2、Inception-v3、Inception-v4和Inception-ResNet。

### GoogLeNet

GoogLeNet采用了一种被称为"Inception"的模块化网络结构，该结构可以并行地处理不同尺度的特征，并在最后进行融合。这种结构的设计使得GoogLeNet在保持较少参数的同时，能够提取丰富的图像特征。

GoogLeNet的整体结构相对复杂，由多个Inception模块组成。每个Inception模块内部由多个不同尺度的卷积层和池化层组成，其中包括1x1、3x3和5x5的卷积操作，以及1x1的池化操作。这种并行的结构允许网络同时学习不同尺度下的特征。

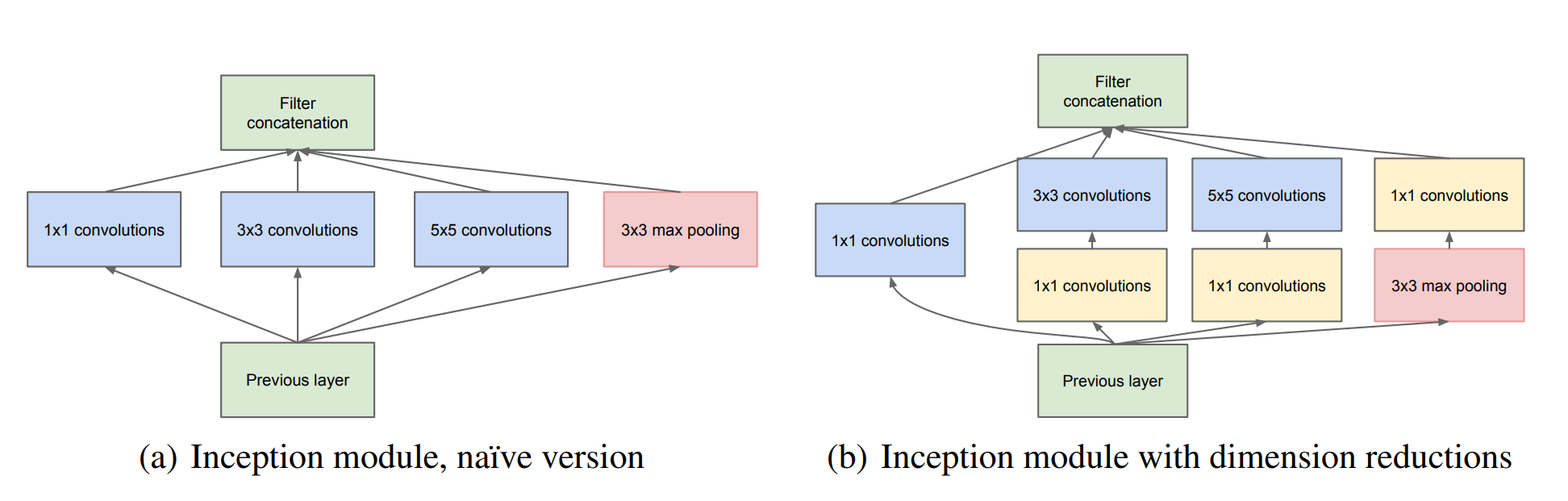


图 4‑2 Inception模块[2]

除了Inception模块外，GoogLeNet还包含了全局平均池化层（Global Average Pooling Layer），以及多个全连接层和输出层。

GoogLeNet主要用于图像分类和物体识别任务。通过Inception模块的设计，GoogLeNet可以在减少参数和计算复杂度的同时，提高图像分类任务的准确性。在2014年的ImageNet图像分类挑战赛上，GoogLeNet取得了显著的成果，获得了较低的错误率，超过了当时其他参赛模型。

GoogLeNet的提出对深度学习和计算机视觉领域具有重要的学术意义。

首先，它引入了"Inception"模块的概念，通过并行处理不同尺度的特征，提高了网络的表达能力。这种模块化的设计思想对后续的网络架构设计产生了深远影响，例如后续的Inception-v2、Inception-v3等。

其次，GoogLeNet证明了通过合理的网络设计和参数优化，可以在减少计算复杂度的同时，保持较高的准确率。这为后续更深、更复杂的网络设计提供了启示，促进了深度学习在计算机视觉领域的进一步发展。

最后，GoogLeNet的成功也加速了深度学习在工业和实际应用中的普及，推动了图像分类、目标检测和图像生成等任务的进展。

### Inception-v2

Inception-v2是对原始的GoogLeNet（Inception-v1）的改进版本之一，由Google团队在2015年提出。它在保留GoogLeNet的"Inception"模块化结构的基础上，引入了一些重要的改进，旨在提高网络的训练速度和准确性。

主要的改进点如下：

1. 批标准化（Batch Normalization）：

Inception-v2引入了批标准化技术，将批量数据在训练过程中进行标准化，以加速训练收敛过程并减少对初始权重的敏感性。批标准化有助于缓解梯度消失和梯度爆炸问题，同时还能够充分利用网络的非线性激活函数。

1. 1x1卷积：

Inception-v2使用了更多的1x1卷积操作。1x1卷积的作用是通过降低通道数来减少计算复杂度，并引入非线性激活函数来增加网络的表达能力。1x1卷积的使用可以对不同通道之间的特征进行线性组合和重组，有助于提取更丰富的特征表示。

1. 网络结构优化：

Inception-v2对Inception模块的结构进行了优化。它引入了一种称为"Inception Module B"的变体，相比于原始的"Inception Module A"具有更高的计算效率和更少的参数量。此外，Inception-v2还通过精心设计的超参数设置，进一步改善了网络的性能和效率。

总的来说，Inception-v2在保留原始GoogLeNet的基本架构和思想的同时，通过批标准化、1x1卷积和结构优化等改进，提高了网络的训练速度和准确性。这些改进使得Inception-v2成为了更加高效和强大的深度卷积神经网络，为图像分类和计算机视觉任务的研究和应用提供了重要的基础。

### Inception-v3

Inception-v3是Google团队在2015年提出的深度卷积神经网络架构，是对之前的Inception-v2进行改进和扩展的版本。它在图像分类和物体识别任务上取得了显著的准确性提升，并成为计算机视觉领域的重要里程碑。

Inception-v3相对于Inception-v2的改进主要集中在以下几个方面：

1. 辅助分类器（Auxiliary Classifiers）：

Inception-v3引入了辅助分类器的概念，这些分类器位于网络的中间层，并通过额外的分类损失函数来辅助网络的训练。这些辅助分类器有助于梯度的传播和减轻梯度消失的问题，使得网络能够更好地学习有效的特征表示。

1. 更深的网络结构：

Inception-v3相比于Inception-v2具有更深的网络结构，它包含了更多的卷积层和Inception模块。通过增加网络的深度，Inception-v3能够提取更复杂、更抽象的特征，并具有更强大的表达能力。

1. 使用较小的卷积核：

Inception-v3采用了更小的卷积核（如3x3），以代替较大的卷积核（如5x5）。这种设计选择可以减少网络的参数量，降低计算复杂度，并提高网络的特征提取能力。

1. 更复杂的Inception模块：

Inception-v3中的Inception模块相比于之前的版本更加复杂。它使用了更多的并行分支，包括1x1、3x3和5x5的卷积操作，以及1x1的池化操作。这种并行结构允许网络同时学习不同尺度下的特征，提高了网络的表达能力。

Inception-v3在2015年的ImageNet图像分类挑战赛上取得了出色的成绩，显示出比之前的版本更高的准确性。它的成功在深度学习和计算机视觉领域产生了重要的影响，激发了更多深层网络的研究和应用。同时，Inception-v3的设计思想和技术手段也为后续网络架构的发展提供了重要的参考和启示。

### Inception-v4

Inception-v4是由Google团队在2016年提出的深度卷积神经网络架构，它是对之前的Inception系列网络的进一步改进和扩展。Inception-v4的目标是提高图像分类和目标检测任务的准确性，并探索更深、更强大的网络结构。

以下是Inception-v4的主要特点和改进：

1. 更深的网络结构：

Inception-v4相比于之前的版本具有更深的网络结构。它通过增加网络的层数和卷积层的深度，提高了网络的表达能力和特征提取能力。更深的网络结构可以捕捉到更丰富和复杂的特征表示，从而提高任务的准确性。

1. 更复杂的Inception模块：

Inception-v4引入了一系列新的Inception模块，包括称为"Inception-ResNet"的模块。这些模块融合了Inception模块和ResNet模块的优点，通过使用ResNet的残差连接来构建更深、更强大的网络。Inception-ResNet模块可以缓解梯度消失问题，加速训练收敛，并提高网络的准确性。

1. 更多的并行分支：

Inception-v4增加了更多的并行分支，使得网络能够同时学习不同尺度和不同复杂度的特征。它使用了多个卷积核尺寸（如1x1、3x3、5x5），并通过1x1卷积和池化操作进行特征融合。这种并行结构增加了网络的灵活性和表达能力，有助于提高任务的性能。

1. 更小的网络尺寸：

Inception-v4通过减少网络的参数量和计算复杂度，使得网络更加轻量化。它使用了较小的卷积核尺寸，以及其他一些技巧来降低网络的参数数量。这使得Inception-v4在计算资源有限的情况下仍能保持较高的准确性。

Inception-v4在多个图像分类和目标检测任务中取得了优秀的性能，显示出比之前版本更高的准确性和鲁棒性。它的设计思想和技术手段对于深度学习和计算机视觉领域的发展具有重要的指导意义，激发了更多深层网络的研究和应用。

## ResNet

ResNet（残差网络）是一种深度卷积神经网络架构，由微软研究院的研究团队于2015年提出。它通过使用残差连接（residual connection）来解决深度网络中的梯度消失和网络难以优化的问题，使得可以训练更深的网络结构。

从AlexNet到VGGNet，再到Inception，网络深度越来越深，模型效果越来越好，这使得当时的研究者们深信“模型越深，准确率越高”。然而，ResNet团队经过实验，随着网络深度不断加深，模型准确率的确在不断提高，可抵达某个点以后，模型准确率就开始毫无征兆地大幅降低。这与之前行业的共识“越深越好”显然是矛盾的，而ResNet团队称其为“退化”（degradation）。他们最终将退化现象归因为深层网络难以实现恒等映射（identity mapping），即y=x这样的简单函数。这乍一听很不可思议，理论上能拟合任何函数的神经网络居然无法实现恒等映射这么简单的变换吗？他们对此的解释是，深层网络在层层转换中加入了太多了非线性变换，这虽然有利于加强数据分类能力，但是也导致网络在非线性的路上越走越远，到了一定深度以后，拟合线性变换的成本都会到无法接受的地步。所以，ResNet团队按照这个思路，加入了残差连接。

残差连接允许信息在网络中更直接地传播。具体来说，它通过将输入直接添加到模块的输出上，使得网络可以学习残差部分，即输入与输出之间的差异。这样，网络可以更轻松地学习恒等映射，并且可以将较深层的网络训练成较浅层网络的近似。

使用残差连接的ResNet模块被称为残差模块（Residual Block）。每个残差模块通常由两个或三个卷积层组成，其中包括一个跳跃连接（skip connection），即将输入直接添加到模块的输出上。这样的设计使得网络能够通过残差路径直接传播梯度，减轻了梯度消失的问题，同时允许网络在训练过程中更好地优化和学习有效的特征表示。

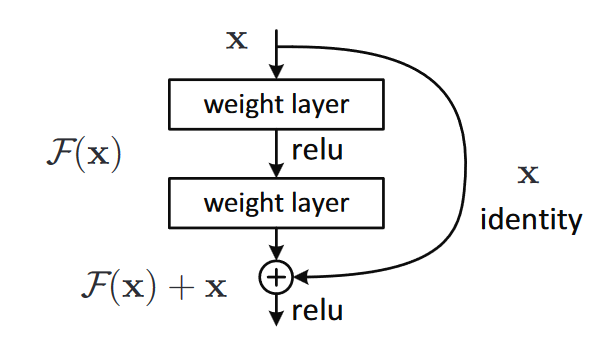


图 4‑3 残差模块[3]

ResNet的论文中提出的网络深达34层。其中大量堆叠了3x3卷积核的卷积层，并且使用残差连接的方式。

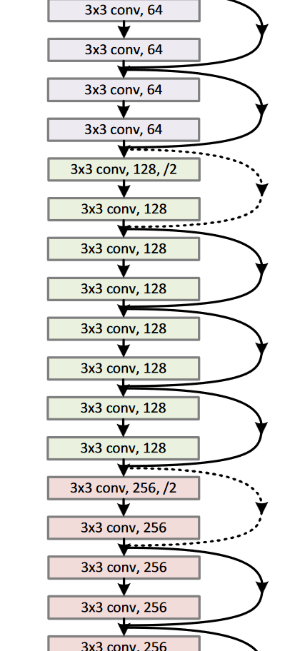


图 4‑4 ResNet部分结构[3]

ResNet的主要创新在于通过残差连接实现了更深的网络结构，极大地推动了深度卷积神经网络的发展。通过堆叠多个残差模块，ResNet可以构建非常深的网络，如ResNet-50、ResNet-101和ResNet-152等，这些网络在图像分类、目标检测和语义分割等计算机视觉任务上取得了卓越的性能。ResNet的设计思想和技术手段对于深度学习和计算机视觉领域的发展产生了深远的影响，激发了更多基于残差连接的网络架构的研究和应用。

## 3x3卷积真的是最好的选择吗？

卷积核作为卷积神经网络的核心组件之一，决定了特征提取的方式和效果。回头观察我们已经了解过的几种CNN，不难发现，我们传统上常使用3x3的卷积核。然而，我们一定要使用3x3卷积核吗？

的确，3x3的卷积核有不少好处。3x3卷积核具有较小的尺寸，减少了网络的参数量，从而降低了计算和存储成本；3x3卷积核可以通过堆叠和组合实现任意大小的感受野，因此它具有较大的灵活性；3x3卷积核还具有对称性和平移不变性的特点，有助于提取图像中的空间特征。然而，尽管3x3卷积核具有多种优势，但也存在其他卷积核尺寸的优势。首先，较大的卷积核（如5x5或7x7）可以覆盖更大的感受野，有助于提取图像中的全局特征。对于较复杂的任务，使用较大的卷积核可能会更加有效。其次，较小的卷积核（如1x1）可以用于降维或增加非线性。1x1卷积核通常用于通道数的调整和特征的融合，同时减少了计算量。

在CVPR2022中甚至提出了一种使用25x25，甚至31x31的超大卷积核的模型——RepLKNet[4]。他们的实验结果表明，在结构重参数化、深度卷积等设计要素的加持下，超大卷积在保持了大感受野的同时，还能够有优秀的计算复杂度，而模型效果在目标检测和语义分割等任务上超过 Swin Transformer，甚至远超传统小卷积模型。这似乎颠覆了我们的想象，他们居然用上了这么大的卷积核，还能算得这么快！而同年的SLaK模型[5]甚至把卷积核提到了61x61的大小，这些最近的工作都证明了大卷积核的可行性。

过去，我们使用小卷积核最重要的一个原因是算力限制。随着卷积核的尺寸增长，计算复杂度将以平方级别地上升，所以我们才更多使用小卷积核。然而，小卷积核的一大缺点是感受野小，这也是一些目标检测任务中小卷积核难以检测到大物体的原因。而如今我们在算力飞速发展的时代，是否可以像GPT一样力大砖飞呢？答案是可以。在下一节中，我们将介绍RepLKNet这样的大卷积核卷积神经网络如何降低计算复杂度。

## 降低计算复杂度的再努力

为了降低大卷积核的计算复杂度，可以使用深度卷积技术。

一般而言，大核卷积的计算量很大，计算复杂度将随卷积核尺寸的增长成平方级地上升。不过，深度卷积则可以大大优化这个问题。

我们称之前介绍的卷积操作为Dense卷积。对于(in\_channel, height, width)形状的输入，设置out\_channel个(in\_channel, kernel\_size, kernel\_size)形状的卷积核，Dense卷积将每个卷积核应用于输入进行计算，将得到形状为(out\_channel, out\_height, out\_width)的特征图。

而深度卷积（depth-wise convolution）的做法是，按每个通道进行卷积。对于(in\_channel, height, width)的输出，卷积核就一定是(in\_channel, kernel\_size, kernel\_size)的形状。然后，逐个通道地进行卷积。因为深度卷积的特点，in\_channel和out\_channel相等，所以最后得到的特征图形状将是(in\_channel, out\_height, out\_width)。

经过一些推理可以得知，深度卷积在参数量上和计算量上仅为Dense卷积的，这就是为什么在大核网络上深度卷积比Dense卷积更好。

## 小结

在本章中，我们介绍了自2012年AlexNet以来的几种重要的卷积网络结构，并对各个网络引入的相关技术进行了说明。我们还讨论了卷积核大小的设置，提出卷积核并不一定要小，大核卷积同样可以有很好的性能，并且介绍了降低大核卷积计算复杂度的深度卷积算法。

## 参考文献

[1] A. Krizhevsky, I. Sutskever和G. E. Hinton, 《ImageNet classification with deep convolutional neural networks》, *Commun. ACM*, 卷 60, 期 6, 页 84–90, 5月 2017, doi: 10.1145/3065386.

[2] C. Szegedy等, 《Going Deeper with Convolutions》. arXiv, 2014年9月16日. 见于: 2023年5月15日. [在线]. 载于: http://arxiv.org/abs/1409.4842

[3] K. He, X. Zhang, S. Ren和J. Sun, 《Deep Residual Learning for Image Recognition》. arXiv, 2015年12月10日. doi: 10.48550/arXiv.1512.03385.

[4] X. Ding, X. Zhang, J. Han和G. Ding, 《Scaling up your kernels to 31x31: Revisiting large kernel design in cnns》, 收入 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 页 11963–11975.

[5] S. Liu等, 《More ConvNets in the 2020s: Scaling up Kernels Beyond 51x51 using Sparsity》. arXiv, 2023年3月3日. doi: 10.48550/arXiv.2207.03620.

[6] 《深度学习算法(二)-CNN发展历程》, 知乎专栏. https://zhuanlan.zhihu.com/p/402450854 (见于 2023年6月3日).

[7] 《AlexNet Architecture: A Complete Guide》. https://kaggle.com/code/blurredmachine/alexnet-architecture-a-complete-guide (见于 2023年6月4日).

[8] 《经典神经网络 | 从Inception v1到Inception v4全解析》, 知乎专栏. https://zhuanlan.zhihu.com/p/194382937 (见于 2023年6月4日).

[9] 《2022\_07\_11\_10\_51x51的kernelsize暴力美学：SLaK论文解读》, 知乎专栏. https://zhuanlan.zhihu.com/p/540426043 (见于 2023年6月4日).

[10] 《凭什么 31x31 大小卷积核的耗时可以和 9x9 卷积差不多？| 文末附 meetup 回顾》, 知乎专栏. https://zhuanlan.zhihu.com/p/479182218 (见于 2023年6月4日).

[11] 《残差神经网络（ResNet）》, 知乎专栏. https://zhuanlan.zhihu.com/p/101332297 (见于 2023年6月4日).